GeoShield: Safeguarding Geolocation Privacy from Vision-Language Models via Adversarial Perturbations

Xinwei Liu[1,2] , **Xiaojun Jia**[3*], **Yuan Xun**[1,2], **Simeng Qin**[4], **Xiaochun Cao**[5*]

[1]Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Nanyang Technological University, Singapore
[4]Northeastern University, China
[5]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, China
{liuxinwei,xunyuan}@iie.ac.cn, jiaxiaojunqaq@gmail.com, qinsimeng@neuq.edu.cn, caoxiaochun@mail.sysu.edu.cn

Abstract

Vision-Language Models (VLMs) such as GPT-4o now demonstrate a remarkable ability to infer users' locations from public shared images, posing a substantial risk to geoprivacy. Although adversarial perturbations offer a potential defense, current methods are ill-suited for this scenario: they often perform poorly on high-resolution images and low perturbation budgets, and may introduce irrelevant semantic content. To address these limitations, we propose ***GeoShield***, a novel adversarial framework designed for robust geoprivacy protection in real-world scenarios. GeoShield comprises three key modules: a feature disentanglement module that separates geographical and non-geographical information, an exposure element identification module that pinpoints geo-revealing regions within an image, and a scale-adaptive enhancement module that jointly optimizes perturbations at both global and local levels to ensure effectiveness across resolutions. Extensive experiments on challenging benchmarks show that GeoShield consistently surpasses prior methods in black-box settings, achieving strong privacy protection with minimal impact on visual or semantic quality. To our knowledge, this work is the first to explore adversarial perturbations for defending against geolocation inference by advanced VLMs, providing a practical solution to escalating privacy concerns.

Code — https://github.com/thinwayliu/Geoshield

Introduction

Recently, Vision-Language Models (VLMs) have emerged as a powerful paradigm that bridges computer vision and natural language processing (Alayrac et al. 2022; Yin et al. 2024; Liu et al. 2023b; Zhu et al. 2023). By jointly modeling visual and textual modalities, VLMs enable a wide range of capabilities, including image captioning (Li et al. 2024; Sarto, Cornia, and Cucchiara 2025), visual question answering (Kuang et al. 2025), and complex multimodal reasoning (Yang et al. 2023). Commercial large-scale VLMs (LVLMs), such as GPT-4, Claude 3.5, and Gemini 2.0, have seen widespread adoption due to their strong performance.

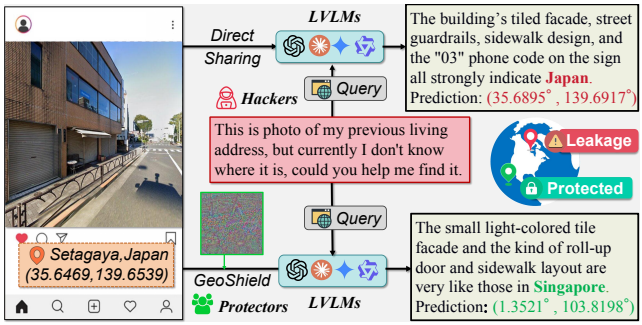
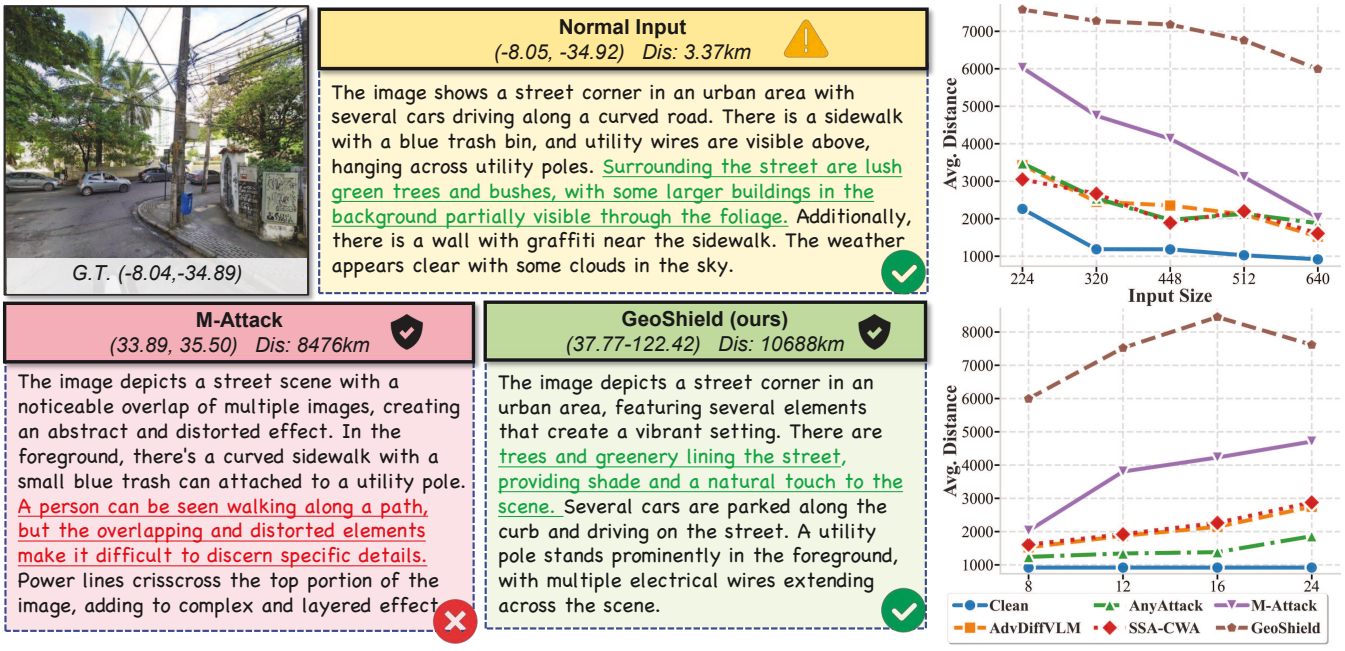


Figure 1: Public image sharing exposes users to geoprivacy threats, as LVLMs can accurately infer locations from visual content. GeoShield applies imperceptible perturbations to disrupt such inference and safeguard user privacy.

However, as the capabilities of VLMs continue to escalate, so do the associated privacy risks (Liang et al. 2022b; enhancing face obfuscation guided by semantic-aware attribution maps 2023; Guo et al. 2023; Dong et al. 2023a; Liang et al. 2024; Gong et al. 2024), particularly concerning geolocation inference (Mendes et al. 2024). Recent studies have highlighted the powerful geolocation abilities of these models (Luo et al. 2025; Zhang et al. 2025b; Jay et al. 2025): VLMs can not only recognize well-known landmarks, but also infer highly accurate geographic coordinates by analyzing subtle visual cues such as lighting conditions, vegetation, and architectural features. This level of inference closely mirrors the mechanics of GeoGuessr, where expert players deduce locations based on minimal information. The advent of VLMs has dramatically lowered the technical barriers for such inferences. For example, a photo casually shared on social media may be collected by malicious actors and queried using advanced VLMs to infer sensitive details, such as a user's home address, workplace, or frequent locations (see Fig. 1). Consequently, safeguarding geographic privacy while preserving the convenience and value of image sharing has become a pressing research challenge.

Recent studies have explored the use of adversarial perturbations (Liang, Wei, and Cao 2021; Liang et al. 2020,

*Correspondence to: Xiaojun Jia and Xiaochun Cao.



(a) VLM responses and predictions on normal and protected inputs.

(b) Impact of input size and budget.

Figure 2: Comparison of semantic consistency and protection effectiveness: (a) M-attack introduces incorrect semantics while ours preserves accurate descriptions; (b) Baseline declines with larger input size and smaller budget, while ours remains stable.

2022c,a; Wang et al. 2023; Liu et al. 2023a; Wang et al. 2025; Wei et al. 2018) as a means of defending against malicious AI models and applications (Van Le et al. 2023; Le et al. 2022; Liu et al. 2022). For geo-privacy protection, a practical strategy is to add carefully crafted adversarial perturbations to public images, thereby preventing unauthorized geolocation inference. However, conducting effective adversarial attacks against advanced commercial models remains challenging due to their closed-source nature. To address this issue, recent work (Zhang, Yi, and Sang 2022; Yin et al. 2023) has shown that integrating multiple white-box visual encoders and minimizing global feature distances between adversarial and target examples can significantly improve the transferability, thus enabling effective attacks against closed-source commercial models (Xu et al. 2024; Lu et al. 2023; Wei et al. 2023).

We systematically evaluated existing adversarial attacks for VLMs (AdvDiffVLM (Guo et al. 2024), AnyAttack (Zhang et al. 2025a), SSA-CWA (Dong et al. 2023b), M-Attack (Li et al. 2025)) for geographic protection. These methods manipulate perturbed image features to align with target images from different locations, misleading models to predict incorrect geolocations. However, our experiments reveal that all approaches face three major challenges.

Firstly, targeted attack methods are fundamentally incompatible with the objective of geo-privacy protection. While these attacks (e.g., M-Attack) aim to mislead VLMs into predicting incorrect locations by aligning image features with those of another image, they don't focus the features and regions within an image that might leak geographical information. Consequently, the generated perturbations often fail to significantly reduce geo-localization accuracy. Moreover, by forcing feature alignment with an unrelated image, these methods not only offer suboptimal privacy protection but also distort the original content and introduce irrelevant semantic information, as illustrated in Fig. 2(a). Such modifications can degrade user experience and undermine other social applications, like content classification for recommendations on social media platforms.

Second, existing attacks typically generate low-resolution perturbations tailored to the input size of visual encoders (e.g., 224×224 for CLIP). Moreover, these methods are primarily evaluated on images with simple backgrounds and few objects, which makes them unsuitable for the high-resolution, object-rich images on social media. As Fig. 2(b) illustrates, upsampling these optimized low-resolution perturbations to higher-resolution images significantly degrades their effectiveness, and this protective effect diminishes even further as image resolution increases.

Third, the effectiveness of these attacks often depends on an excessively high perturbation budget (e.g., 16/255), which substantially degrades image quality. Fig. 2(b) further demonstrates that when the perturbation budget is constrained, existing methods generally fail to provide adequate geographic privacy protection.

To address these challenges, we propose GeoShield, a novel perturbation generation framework for real-world geo-privacy protection. GeoShield is designed to produce visually imperceptible yet highly effective perturbations that disrupt the geolocation capabilities of VLMs while preserving semantic integrity. It consists of three key modules: (1) Geographical and Non-Geographical Feature Disentanglement

(GNFD), which leverages VLMs to produce generic image descriptions and disentangle geographical features from general semantic features; (2) Geographical Exposure Element Identification (Geo-EE), which localizes geographical exposure elements (e.g., landmarks, architecture) using a combination of VLMs and object detection; and (3) Perturbation Scale Adaptive Enhancement (PSAE), which jointly optimizes perturbations over global and local patches to ensure effectiveness across varying image resolutions. These modules suppress geo-relevant features while preserving alignment with non-geographic semantics. Extensive experiments show that our method consistently outperforms existing methods under black-box settings.

Our contributions can be summarized as follows:

- We are the first to leverage adversarial perturbation to protect user geolocation privacy against powerful VLMs.
- We conduct a systematic evaluation of existing adversarial methods in the context of geo-privacy and reveal their limitations under realistic scenarios.
- We propose GeoShield, a novel framework that disentangles geo-relevant features, localizes geo-exposing regions, and enhances perturbation robustness across scale.
- Extensive experiments show that GeoShield consistently outperforms existing baselines under black-box conditions, achieving strong privacy protection with minimal semantic or visual degradation.

Related Work

Image Geo-Localization

Geolocation inference refers to the ability to determine precise geographic coordinates (latitude and longitude) from one or more input images (Clark et al. 2023; Vivanco Cepeda, Nayak, and Shah 2023). Traditionally, common localization approaches have relied on image-to-image retrieval techniques (Suresh, Chodosh, and Abello 2018; Wu and Huang 2022; Berton, Masone, and Caputo 2022). However, a significant limitation of these methods is the prohibitive requirement for large-scale global reference datasets, which renders them impractical for broad application. Another approach involves classification-based methods, where geographical maps are partitioned into discrete categories and models are trained to classify images into these predetermined regions (Theiner, Müller-Budack, and Ewerth 2022; Haas et al. 2024). Nevertheless, the generalization capabilities of these approaches remain constrained by fixed geographic granularity and the need for extensive annotated datasets tailored to each specific region.

Recent advancements in VLMs have demonstrated an unexpected proficiency in predicting geographic locations, despite not being explicitly trained for geolocation tasks (Wang et al. 2024; Zhang et al. 2024; Yang et al. 2024). Notably, Jay et al. (2025) conducted an evaluation of various open-source and closed-source VLMs for their precise geolocation capabilities, revealing surprisingly high accuracy. Furthermore, Luo et al. (2025) performed a study on the potential privacy risks associated with the visual reasoning capabilities of these models.

Adversarial Attacks on VLMs

Adversarial attacks on VLMs aim to induce incorrect model outputs by adding imperceptible perturbations (Jia et al. 2025). Given that many commercial LVLMs are closed-source, black-box attacks—especially transfer-based attacks—are more practical. These transfer-based attacks generate adversarial examples on surrogate models such as CLIP (Radford et al. 2021) and BLIP (Li et al. 2022), which are then successfully transferred to target models. AttackVLM (Zhao et al. 2023) was the first to introduce this strategy, demonstrating that image-to-image feature matching achieves better transferability than image-to-text optimization. Subsequent approaches like CWA (Chen et al. 2023) and SSA-CWA (Dong et al. 2023b) have further enhanced transferability by leveraging ensemble surrogates and frequency-based transformations, showing notable success against commercial LVLMs such as Google Bard. Additionally, methods including AnyAttack (Zhang et al. 2025a) and AdvDiffVLM (Guo et al. 2024) incorporate self-supervised pretraining and diffusion guidance to generate transferable adversarial examples, albeit often at the expense of image quality or increased complexity. M-Attack (Li et al. 2025) further improves transfer success by introducing random cropping and resizing during optimization.

Methodology

Preliminary

Given a test geographical dataset $D = \{(I_n, G_n)\}_{n=1}^{N}$, where I_n denotes the n-th image and $G_n = (\phi_n, \lambda_n)$ represents its true geographical coordinates (latitude and longitude), our core objective is to generate an imperceptible perturbation δ_n for each image I_n. The resulting protected image $I'_n = I_n + \delta_n$ is designed to mislead a target VLM, denoted as f_t, into predicting incorrect coordinates $G'_n \neq G_n$, thereby providing geographical privacy protection for users.

To quantify the effectiveness of protection, we compute the geolocation error as the great-circle distance (in kilometers) between the predicted and true coordinates. This error is calculated using the Haversine formula, which estimates the shortest distance over the Earth's surface between two points. Given two locations with coordinates (ϕ_1, λ_1) and (ϕ_2, λ_2), the distance d is computed as:

$$d = R \cdot \arctan 2\left(\sqrt{\mathrm{hav}(\theta)}, \sqrt{1-\mathrm{hav}(\theta)}\right), \quad (1)$$

where R is the Earth's mean radius, and θ is the central angle between the two points. The haversine of θ is defined as:

$$\mathrm{hav}(\theta) = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2\left(\frac{\Delta\lambda}{2}\right). \quad (2)$$

Our objective is to maximize the geographical distance between the predicted coordinates G'_n and the ground-truth coordinates G_n. This can be formulated as a constrained optimization problem:

$$\max_{\delta_n} \quad d\big(f_t(I_n+\delta_n), G_n\big) \quad \text{s.t.} \quad \|\delta_n\|_\infty \leq \epsilon, \quad (3)$$

where ϵ denotes the perturbation budget, constraining the magnitude of adversarial noise under the ℓ_∞ norm.

In this work, we consider a black-box setting, where the protector has no access to the internal architecture, parameters, or training data of the target VLM. This aligns with realistic deployment scenarios, as privacy defenses are typically applied before the image is exposed to potential hackers. Moreover, commercial LVLMs such as GPT-4o, Claude-3.5, and Gemini-2.5 are accessible only via APIs, making white-box, gradient-based geolocation attacks impractical.

Limitations of Existing Baselines

Recent studies have shown that adversarial examples crafted using an ensemble of pre-trained image encoders exhibit significantly improved transferability and can successfully perform targeted attacks against commercial VLMs. Motivated by these findings, we adopt an ensemble-based strategy for geographic privacy protection. In the remainder of this paper, we assume both visual and textual encoders are implemented as ensembles of paired encoders, with each pair operating in a shared feature space. Given a protecting image x and a target image x_t sampled from a geographically distant location, our objective is to generate a targeted perturbation δ such that the visual features of the perturbed image $x+\delta$ are closely aligned with those of x_t. This can be formulated as the following constrained optimization problem:

$$\min_{\delta} \sum_{i=1}^{M} \left[\mathcal{S}(f_{\theta_i}(x+\delta), f_{\theta_i}(x_t))\right] \quad \text{s.t.} \quad \|\delta\|_\infty \leq \epsilon, \quad (4)$$

where $f_{\theta_i}(\cdot)$ denotes the i-th image encoder in the ensemble and $\mathcal{S}(\cdot,\cdot)$ is a feature-space similarity loss (e.g., cosine distance). However, applying existing attack baselines to solve the above problem encounters three significant challenges:

1. **Inconsistent Objective:** The primary goal of baseline methods is typically to mislead the model into classifying the perturbed image into a specific object or content. This differs from the goal of geographical privacy protection, which is to induce incorrect location predictions. As a result, these perturbations may not effectively reduce geolocation accuracy. Even though M-Attack reduces accuracy, it often severely distorts semantic features and introduces incorrect descriptions, as shown in Fig. 2 (a), thus compromising the usability of protected images in downstream applications.
2. **Low-Resolution Perturbations:** Most existing attack methods generate perturbations for low-resolution inputs (e.g., 224×224 for CLIP), and are evaluated on images with simple backgrounds and few objects. In practice, user-uploaded social media images are typically high-resolution and object-rich. Applying low-resolution noise to such images via upsampling will significantly reduce perturbation effectiveness. Fig. 2 (b) shows that baseline effectiveness declines rapidly as input size increases, especially for M-Attack.
3. **Excessive Budget:** To improve attack performance, baselines often adopt large perturbation budgets (e.g., $16/255$), leading to noticeable image quality degradation and poor user acceptance. However, as shown in Fig. 2 (b), baselines generally fail to maintain privacy protection under more realistic, lower-budget constraints.

Our Proposed Method: GeoShield

To safeguard geo-privacy in high-resolution images while maintaining semantic integrity on a smaller perturbation budget, we introduce GeoShield, a novel framework comprises three core modules. The overall architecture of GeoShield is illustrated in Fig 3.

Geographical and Non-Geographical Feature Disentanglement (GNFD) To effectively prevent VLMs from accurately inferring the geographic location from an image, it is essential to identify and suppress those directions in the visual features that encode geolocation information. At the same time, retaining features that are unrelated to geography ensures the preservation of the original semantic content. Here we introduce a feature decoupling mechanism. Specifically, we assume that an image representation $\mathbf{z}$, extracted by an ensemble of pre-trained image encoders, can be decomposed into two components: a geography-specific vector $\mathbf{z}_{geo}$ and a non-geographic semantic vector $\mathbf{z}_{non\text{-}geo}$, formally expressed as

$$\mathbf{z} = \mathbf{z}_{\text{geo}} + \mathbf{z}_{\text{non-geo}}. \quad (5)$$

However, precisely disentangling geographical and non-geographical features in the feature space remains challenging, as most feature representations are inherently entangled and lack explicit annotations separating them. Thus, we employ an auxiliary VLM to implicitly approximate these components. Specifically, we design a tailored prompt for a powerful VLM (e.g., GPT-4o) to generate a detailed non-geographical textual description $T_{\text{non-geo}}$ of the original image x, explicitly excluding any geographical clues such as place names, landmarks, or city/country identifiers (the exact prompt is provided in the appendix). This geo-filtered description is intended to capture the general semantic content of the image while omitting geographic information. We then encode $T_{\text{non-geo}}$ using an ensemble of textual encoders $g_{\theta_i}(\cdot)$, aligned with those used for image feature extraction, to obtain the textual feature:

$$\mathbf{z}_{\text{non-geo}} \approx g_{\theta_i}(T_{\text{non-geo}}). \quad (6)$$

We regard $\mathbf{z}_{\text{non-geo}}$ as the non-geographical feature component and use it to estimate the geographical features. Given an protecting image x, its visual feature can be denoted as $f_{\theta_i}(x)$. Therefore, we can approximate the geographical feature vector by subtracting $\mathbf{z}_{\text{non-geo}}$ from the original feature. Formally, the geographical component $\mathbf{z}_{\text{geo}}$ is given by:

$$\begin{aligned} \mathbf{z}_{\text{geo}} &= \mathbf{z} - \mathbf{z}_{\text{non-geo}} \\ &\approx f_{\theta_i}(x) - g_{\theta_i}(T_{\text{non-geo}}). \end{aligned} \quad (7)$$

Geographical Exposure Element Identification (Geo-EE) Disentangling geographical features solely at the global level is often insufficient to capture all the cues VLMs use for location prediction. Moreover, many reasoning-based models, such as o3, perform local recognition across different image regions before geographic localization. This highlights the need to identify local regions or visual elements that could expose geographical information.

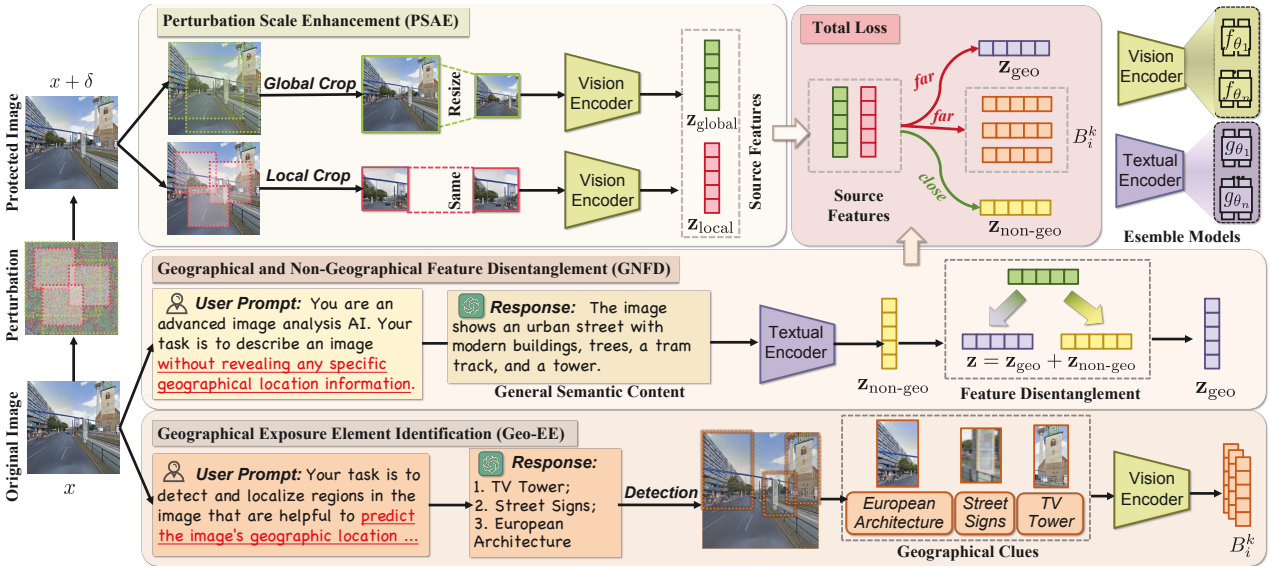


Figure 3: Overview of the GeoShield framework. GeoShield consists of three modules (GNFD, Geo-EE, and PSAE) that collaboratively suppress geographical cues while preserving semantic integrity in high-resolution images.

To address this, we introduce the Geographical Exposure Element Identification (Geo-EE) module. We again employ an auxiliary VLM to identify and generate the names of objects or landmarks within the image that may reveal geographic information, such as "European Architecture" or "TV Tower." These entities are assumed to be strongly associated with specific locations. We then use these names as prompts for a pre-trained object detection model (e.g., GroundingDINO (Liu et al. 2023c) or SAM (Kirillov et al. 2023)) to identify and output a set of local geo-indicative bounding boxes, denoted as $\mathcal{B} = \{b_1, b_2, \ldots, b_K\}$. For each box b_k, we crop the corresponding image region x_{b_k} and extract visual features using the same ensemble of image encoders $f_{\theta_i}(\cdot)$ as in the GNFD module. Formally, the feature extracted from the i-th encoder for the k-th bounding box is:

$$B_i^k = f_{\theta_i}(x_{b_k}) \tag{8}$$

These local features F_{b_k} approximate subsets of the geographical information present in the original image. By isolating such fine-grained cues, we enable more comprehensive protection of geographical privacy.

Perturbation Scale Enhancement Through the GNFD and Geo-EE modules, we extract approximate geographical features ($\mathbf{z}_{\text{geo}}$, F_{b_k}) and non-geographical features ($\mathbf{z}_{\text{non-geo}}$) from an image. However, as previously discussed, perturbations often struggle with scale adaptability on high-resolution images, leading to diminished privacy protection.

Therefore, we propose the Perturbation Scale Adaptive Enhancement (PSAE) module, which employs a joint global and local optimization. We first follow the data augmentation strategy from M-Attack by applying random cropping to the entire image before encoding, which has been shown to significantly improve transferability. Specifically, in each iteration, we perform random cropping on the entire image x to obtain global source features f_{global}. In addition, we simultaneously reinforce perturbations in randomly sampled local regions, each matching the input size of the visual encoders (for example, 224×224 for $x_{\text{patch},t}$), and encode these regions to yield a set of local source features $\{f_{\text{local},t}\}_{t=1}^{N_{\text{patch}}}$. We aggregate these local features by averaging over all sampled patches, resulting in the following formulation:

$$f_{\theta_i}^{\text{local}} = \frac{1}{N_{\text{patch}}} \sum_{t=1}^{N_{\text{patch}}} f_{\theta_i}(x_{\text{patch},t}). \tag{9}$$

By jointly optimizing both f_{global} and $\{f_{\text{local},j}\}$ in a multi-scale manner, PSAE preserves fine-grained details and enables locally refined updates based on the global perturbation. In particular, to further enhance transferability through increased randomness, we use the global source features f_{global} obtained in each iteration as the decomposition targets $f_{\theta_i}(x)$ in the GNFD module.

Total Loss Function: Based on the extracted features, we construct the following loss. The primary objective is to minimize the similarity between the source global and local features of the perturbed image and the geographical features, while maximizing similarity with non-geographical semantic features to preserve semantic integrity. This objective can be formalized as the following problem:

$$\begin{aligned}\min_{\delta} \sum_{i=1}^{N} \Big\{ & \big[\mathcal{S}(f_{\theta_i}^{\text{global}}(x'), \mathbf{z}_{\text{geo}}) + \mathcal{S}(f_{\theta_i}^{\text{local}}(x'), \mathbf{z}_{\text{geo}})\big] \\ & + \alpha \sum_{k=1}^{K} \big[\mathcal{S}(f_{\theta_i}^{\text{global}}(x'), B_i^k) + \mathcal{S}(f_{\theta_i}^{\text{local}}(x'), B_i^k)\big] \\ & - \beta \big[\mathcal{S}(f_{\theta_i}^{\text{global}}(x'), \mathbf{z}_{\text{non-geo}}) + \mathcal{S}(f_{\theta_i}^{\text{local}}(x'), \mathbf{z}_{\text{non-geo}})\big] \Big\} \\ \text{s.t.} \quad & \|\delta\|_\infty \le \epsilon\end{aligned}$$

Dataset	Model	GPT-4o					GPT-4.1					Claude-3.5					Gemini-2.5				
		1km	25km	200km	750km	2500km	1km	25km	200km	750km	2500km	1km	25km	200km	750km	2500km	1km	25km	200km	750km	2500km
Google Street View	Clean	7.3	17.7	41.6	73.4	90.6	9.1	23.2	48.5	78.0	92.8	4.9	9.0	27.5	57.8	78.1	8.7	21.9	47.6	79.2	93.5
	RN	6.9	16.1	40.6	73.2	90.7	9.6	23.7	48.8	77.5	92.3	4.7	8.9	27.2	56.4	77.8	8.9	21.2	47.1	80.3	92.5
	AdvDiffVLM	5.5	13.4	34.2	68.0	87.0	7.7	20.0	43.4	74.4	90.3	2.7	8.4	24.8	53.9	73.8	8.1	20.4	45.6	77.8	93.0
	AnyAttack	6.2	15.9	37.7	68.7	87.3	7.9	20.4	44.4	73.1	89.5	2.8	8.4	24.1	51.4	74.2	8.0	19.8	44.2	77.3	91.2
	SSA-CWA	4.7	12.1	31.8	62.9	83.4	7.2	18.4	40.8	70.0	88.3	1.4	6.1	18.3	44.9	62.1	7.2	18.8	40.6	74.5	90.5
	M-Attack	3.3	9.1	24.5	48.7	71.1	4.9	12.6	30.3	54.9	76.1	1.4	5.3	15.3	35.8	56.3	6.1	16.7	37.6	66.8	86.3
	GeoShield	**1.1**	**2.9**	**7.6**	**17.5**	**33.8**	**1.4**	**3.6**	**9.1**	**20.9**	**37.9**	**0.1**	**1.1**	**5.2**	**12.5**	**27.4**	**4.7**	**13.0**	**32.4**	**59.1**	**80.9**
Img2gps	Clean	14.4	38.9	55.8	71.4	84.6	18.2	46.3	59.4	73.9	87.4	9.1	30.0	43.4	61.9	77.1	18.2	45.5	59.7	74.9	86.4
	RN	14.1	38.8	55.1	70.7	84.7	17.8	44.9	58.9	73.3	85.6	8.6	28.9	42.1	60.5	76.9	18.0	45.4	58.9	73.2	86.0
	AdvDiffVLM	13.8	35.7	49.9	67.4	82.3	17.3	43.5	56.8	70.0	83.1	8.3	26.7	41.4	58.2	74.5	17.8	43.2	56.8	70.1	82.1
	AnyAttack	14.0	36.9	52.1	67.1	81.4	17.4	43.3	56.6	71.5	84.3	8.2	26.5	40.0	57.8	74.3	17.8	40.8	56.5	69.4	81.6
	SSA-CWA	13.5	33.1	47.2	64.7	76.8	15.9	39.8	53.0	67.2	81.3	7.4	24.3	35.1	52.0	69.3	16.1	39.6	53.6	68.5	78.4
	M-Attack	9.2	23.0	32.9	46.2	61.1	13.2	30.1	40.0	50.5	65.3	5.6	16.9	24.1	36.4	52.8	15.2	36.6	49.3	61.6	77.3
	GeoShield	**4.0**	**9.7**	**12.3**	**20.0**	**5.8**	**9.2**	**13.2**	**16.2**	**23.2**	**38.3**	**2.4**	**6.7**	**8.6**	**14.6**	**30.1**	**10.4**	**25.4**	**31.8**	**41.7**	**55.3**

Table 1: Geolocation prediction accuracy (%) at multiple distance levels on Google Street View and Im2GPS3k datasets. GeoShield consistently provides superior geoprivacy protection compared to existing baselines across all black-box VLMs.

where $x' = x + \delta$ is the perturbed image, $f_{\theta_i}(\cdot)$ denotes the i-th visual encoder in the ensemble, and $\mathcal{S}(\cdot, \cdot)$ indicates cosine similarity. α and β are weighting coefficients.

This optimization can be addressed using standard adversarial frameworks such as I-FGSM (Kurakin, Goodfellow, and Bengio 2018), PGD (Madry et al. 2017), or C&W (Carlini and Wagner 2017). Following M-Attack, we adopt a uniformly weighted ensemble with I-FGSM. Full algorithmic details are provided in the appendix.

Experiment

Experimental Settings

Datasets. We conducted experiments on two public geographic image datasets: Google Street View and Im2GPS3k, both of which provide images paired with GPS coordinates. The Google Street View dataset contains 1,602 images from 1,563 unique cities across 88 countries. The Im2GPS3k dataset includes approximately 3,000 geotagged images from sources such as Flickr. Unless otherwise specified, all images were resized to 640×640 pixels. In addition, target images for baseline methods were randomly selected from MSCOCO dataset (Lin et al. 2014).

Implementation Settings. For fair comparison and consistency with prior work, we used three CLIP variants (ViT-B/16, ViT-B/32, and ViT-g-14-laion2B-s12B-b42K) as surrogate models to generate perturbations. The budget was set to $8/255$ under the ℓ_∞ norm to avoid visual degradation, with an attack step size of $1/255$ and 200 attack iterations. Main results are reported on four popular black-box VLMs: GPT-4o, GPT-4.1, Claude-3.5, and Gemini-2.5; Unless otherwise specified, we default to using GPT-4o as both the auxiliary and target VLM for all experiments, and they were conducted on four NVIDIA A100 GPUs (80GB).

Evaluation Metrics. Geolocation accuracy was measured by the Haversine distance between predicted and ground truth coordinates, evaluated at five granularities: street (1 km), city (25 km), region (200 km), country (750 km), and continent (2,500 km). For some experiments, average distance was also reported. To assess semantic preservation, we used VLMs to generate textual descriptions for both original and perturbed images, and measured semantic similarity using BLEU, ROUGE, and BERTScore (BERT-S).

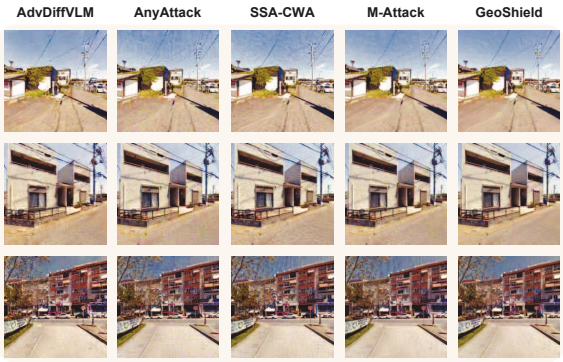


Figure 4: Visualization of different protected images.

Budget ϵ	Method	Llava-1.5			GPT-4o		
		BLEU	ROUGE	BERT-S	BLEU	ROUGE	BERT-S
8/255	AdvDiffVLM	0.50	0.62	0.95	0.14	0.31	0.90
	AnyAttack	0.48	0.59	0.94	0.13	0.31	0.90
	SSA-CWA	0.45	0.57	0.93	0.13	0.31	0.90
	M-Attack	0.22	0.39	0.91	0.09	0.26	0.88
	GeoShield	0.25	0.41	0.92	0.11	0.29	0.89
16/255	AdvDiffVLM	0.48	0.59	0.94	0.12	0.29	0.89
	AnyAttack	0.43	0.55	0.94	0.12	0.30	0.90
	SSA-CWA	0.32	0.47	0.92	0.11	0.29	0.89
	M-Attack	0.15	0.34	0.89	0.07	0.24	0.87
	GeoShield	0.20	0.37	0.91	0.09	0.26	0.89

Table 2: Semantic similarity metrics between original and protected images evaluated on Llava-1.5 and GPT-4o.

Comparative Results

Effectiveness. We evaluate the protection effectiveness of GeoShield by measuring the geolocation accuracy of four commercial VLMs. In Tab. 1, GeoShield consistently achieves the lowest localization accuracy across all models and datasets, significantly outperforming baselines such as M-Attack and SSA-CWA. For instance, on the Google Street View dataset, GeoShield reduces the 1 km-level accuracy from 7.3% (clean) to 1.1% on GPT-4o, and from 4.9% to just 0.1% on Claude-3.5. These confirm effectiveness and transferability of GeoShield under black-box settings.

Fig. 2(b) further illustrates the impact of input resolution and perturbation budget on protection performance. As input size increases from 224 to 640, the effectiveness of base-

Metric	Effectiveness (Avg. Dis) GPT-4o	GPT-4.1	O1	Semantic Consistency BLEU	ROUGE	BERT-S
w/o Geo-EE	7046	6578	5840	0.1067	0.2855	0.8937
w/o GNFD	4481	4229	4307	0.1026	0.2915	0.8945
w/o PSAE	4932	4261	4629	0.1071	0.2918	0.8954
All losses	**7564**	**6868**	**6780**	**0.1078**	**0.2923**	**0.8986**

Table 3: Ablation results for GeoShield on geoprivacy protection and semantic consistency. Each module is essential for strong protection and content preservation.

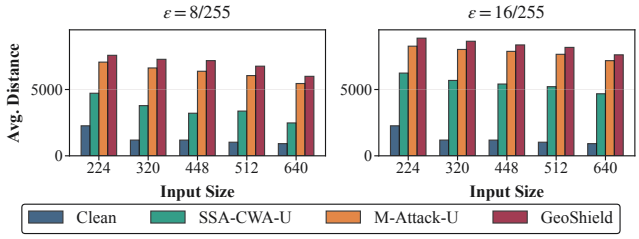


Figure 5: Average distance under untargeted attacks for two perturbation budgets.

line methods drops sharply, whereas GeoShield consistently maintains high protection efficacy. Moreover, ours remains robust even under smaller perturbation budgets $\epsilon = 8$, underscoring its practicality for real-world scenarios.

Semantic Preservation. We further evaluate the semantic consistency between the original and protected images, as shown in Table 2. While GeoShield does not always achieve the highest semantic consistency among all methods, previous effectiveness experiments indicate that, except for M-Attack, other baselines fail to provide an effective protection. For a fair comparison, we argue that a good protection method must first ensure geoprivacy effectiveness before considering semantic consistency. Focusing on the most competitive baseline, we observe that GeoShield consistently achieves higher semantic consistency scores than M-Attack. This illustrates that our perturbations can protect geoprivacy without sacrificing the core semantic content of the image, which is also consistent with the qualitative examples in Fig. 2 (a). Even under a stronger perturbation budget, GeoShield maintains better semantic preservation than M-Attack, supporting practical application for safe image sharing on social media.

Visualization. Fig. 4 shows examples of perturbed images from GeoShield and baselines. GeoShield maintains high visual quality with less artifacts, while methods like AnyAttack and SSA-CWA introduce more visible noise. Moreover, unlike methods that align perturbations to target image, ours avoids semantically misleading distortions, thus better maintaining content authenticity. Additional visualizations and perturbation heatmaps are provided in the appendix.

Discussion

Ablation Study. An ablation study (Tab. 3) demonstrates that removing any GeoShield module results in noticeable performance degradation. Disabling Geo-EE significantly weakens the framework's ability to localize geosensitive regions, leading to reduced protection, particularly for reasoning-based models like o1. Excluding PSAE also causes a marked decline in effectiveness, which results the robustness across varying image resolutions. Most notably, omitting GNFD leads to the greatest decrease in both geoprivacy protection and semantic consistency, underscoring the central role of feature disentanglement. These findings confirm that all three modules are indispensable for robust and reliable geoprivacy protection.

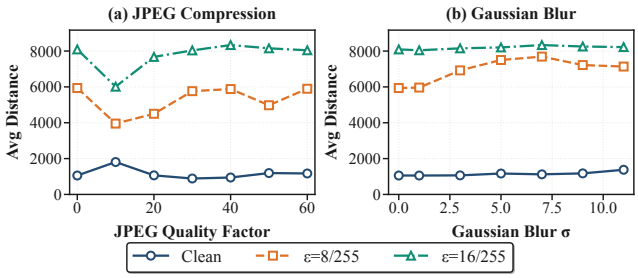


Figure 6: GeoShield maintains geoprivacy protection across varying levels of JPEG compression and Gaussian blur.

Untargeted Attack Baselines. Previous sections primarily presented results for baselines under targeted attack settings. Here, we further evaluate the untargeted attack performance of SSA-CWA and M-Attack (denoted as SSA-CWA-U and M-Attack-U). As shown in Fig. 5, M-Attack-U achieves better protection than SSA-CWA-U, but its effectiveness consistently lags behind that of GeoShield across both perturbation budgets. We speculate that the effectiveness of M-Attack-U may be due to its loss to push features away from the original image representation, which inadvertently suppresses geographical cues, which is partially consistent with ours. Overall, GeoShield provides the most robust geoprivacy protection across all evaluated settings.

Robustness to Transformation. We evaluate the robustness of GeoShield against common image transformations, specifically JPEG compression and Gaussian blur, which frequently occur in real-world image sharing. As shown in Fig. 6, GeoShield consistently provides strong geoprivacy protection across a broad range of JPEG quality factors and blur radii. Even under heavy compression or blurring, the geolocation error remains substantially higher than the clean baseline. These results demonstrate the practical robustness of GeoShield for real-world deployment.

Conclusion

We presented GeoShield, a novel framework for protecting geolocation privacy in VLMs. Extensive experiments demonstrate that GeoShield effectively prevents accurate geolocation inference while preserving both semantic integrity and image usability. Its robust and scalable design makes it a practical solution for applications where geoprivacy is critical. Our work not only offers a promising direction for geographic privacy defense, but also provides insights that can be extended to broader privacy protection tasks.

Acknowledgments

This research is supported by Shenzhen Science and Technology Program (No.KQTD20221101093559018); by the National Natural Science Foundation of China (No.62025604); by Ningbo Science and Technology Innovation 2025 Major Project (2025Z027); by the Open Topics from the Lion Rock Labs of Cyberspace Security (under the project #LRL24009); by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG4-GC-2023-008-1B); by the National Research Foundation Singapore and the Cyber Security Agency under the National Cybersecurity R&D Programme (NCRP25-P04-TAICeN); and by the Prime Minister's Office, Singapore under the Campus for Research Excellence and Technological Enterprise (CREATE) Programme. Any opinions, findings and conclusions, or recommendations expressed in these materials are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore, Cyber Security Agency of Singapore, Singapore.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736.

Berton, G.; Masone, C.; and Caputo, B. 2022. Rethinking visual geo-localization for large-scale applications. In *CVPR*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chen, H.; Zhang, Y.; Dong, Y.; Yang, X.; Su, H.; and Zhu, J. 2023. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*.

Clark, B.; Kerrigan, A.; Kulkarni, P. P.; Cepeda, V. V.; and Shah, M. 2023. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *CVPR*.

Dong, X.; Wang, R.; Liang, S.; Liu, A.; and Jing, L. 2023a. Face Encryption via Frequency-Restricted Identity-Agnostic Attacks. In *ACM MM*.

Dong, Y.; Chen, H.; Chen, J.; Fang, Z.; Yang, X.; Zhang, Y.; Tian, Y.; Su, H.; and Zhu, J. 2023b. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*.

enhancing face obfuscation guided by semantic-aware attribution maps, P. 2023. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *TIFS*.

Gong, J.; Cai, W.; Liang, S.; Guan, Z.; Wang, T.; and Chang, E.-C. 2024. WFCAT: Augmenting Website Fingerprinting with Channel-wise Attention on Timing Features. *arXiv preprint arXiv:2412.11487*.

Guo, J.; Zheng, X.; Liu, A.; Liang, S.; Xiao, Y.; Wu, Y.; and Liu, X. 2023. Isolation and Induction: Training Robust Deep Neural Networks against Model Stealing Attacks. In *ACM MM*.

Guo, Q.; Pang, S.; Jia, X.; Liu, Y.; and Guo, Q. 2024. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *TIFS*.

Haas, L.; Skreta, M.; Alberti, S.; and Finn, C. 2024. Pigeon: Predicting image geolocations. In *CVPR*.

Jay, N.; Nguyen, H. M.; Hoang, T. D.; and Haimes, J. 2025. Evaluating precise geolocation inference capabilities of vision language models. *arXiv preprint arXiv:2502.14412*.

Jia, X.; Gao, S.; Qin, S.; Pang, T.; Du, C.; Huang, Y.; Li, X.; Li, Y.; Li, B.; and Liu, Y. 2025. Adversarial Attacks against Closed-Source MLLMs via Feature Optimal Alignment. *arXiv preprint arXiv:2505.21494*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Kuang, J.; Shen, Y.; Xie, J.; Luo, H.; Xu, Z.; Li, R.; Li, Y.; Cheng, X.; Lin, X.; and Han, Y. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8): 1–36.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Le, T.-N.; Gu, T.; Nguyen, H. H.; and Echizen, I. 2022. Rethinking Adversarial Examples for Location Privacy Protection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, W.; Fan, H.; Wong, Y.; Yang, Y.; and Kankanhalli, M. 2024. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*.

Li, Z.; Zhao, X.; Wu, D.-D.; Cui, J.; and Shen, Z. 2025. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*.

Liang, S.; Gong, J.; Fang, T.; Liu, A.; Wang, T.; Liu, X.; Cao, X.; Tao, D.; and Ee-Chien, C. 2024. Red Pill and Blue Pill: Controllable Website Fingerprinting Defense via Dynamic Backdoor Learning. *arXiv preprint arXiv:2412.11471*.

Liang, S.; Li, L.; Fan, Y.; Jia, X.; Li, J.; Wu, B.; and Cao, X. 2022a. A large-scale multiple-objective method for black-box attack against object detection. In *ECCV*.

Liang, S.; Liu, A.; Liang, J.; Li, L.; Bai, Y.; and Cao, X. 2022b. Imitated detectors: Stealing knowledge of black-box object detectors. In *ACM MM*.

Liang, S.; Wei, X.; and Cao, X. 2021. Generate more imperceptible adversarial examples for object detection. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Liang, S.; Wei, X.; Yao, S.; and Cao, X. 2020. Efficient adversarial attacks for visual object tracking. In *ECCV*.

Liang, S.; Wu, B.; Fan, Y.; Wei, X.; and Cao, X. 2022c. Parallel rectangle flip attack: A query-based black-box attack against object detection. *arXiv preprint arXiv:2201.08970*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, A.; Guo, J.; Wang, J.; Liang, S.; Tao, R.; Zhou, W.; Liu, C.; Liu, X.; and Tao, D. 2023a. {X-Adv}: Physical adversarial object attacks against x-ray prohibited item detection. In *32nd USENIX Security Symposium (USENIX Security 23)*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *NeurIPS*, 36: 34892–34916.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, X.; Liu, J.; Bai, Y.; Gu, J.; Chen, T.; Jia, X.; and Cao, X. 2022. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *ECCV*. Springer.

Lu, D.; Wang, Z.; Wang, T.; Guan, W.; Gao, H.; and Zheng, F. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *ICCV*.

Luo, W.; Lu, T.; Zhang, Q.; Liu, X.; Hu, B.; Zhao, Y.; Zhao, J.; Gao, S.; McDaniel, P.; Xiang, Z.; et al. 2025. Doxing via the Lens: Revealing Location-related Privacy Leakage on Multi-modal Large Reasoning Models. *arXiv preprint arXiv:2504.19373*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mendes, E.; Chen, Y.; Hays, J.; Das, S.; Xu, W.; and Ritter, A. 2024. Granular privacy control for geolocation with vision language models. *arXiv preprint arXiv:2407.04952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Sarto, S.; Cornia, M.; and Cucchiara, R. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*.

Suresh, S.; Chodosh, N.; and Abello, M. 2018. Deepgeo: Photo localization with deep neural network. *arXiv preprint arXiv:1810.03077*.

Theiner, J.; Müller-Budack, E.; and Ewerth, R. 2022. Interpretable semantic photo geolocation. In *WACV*.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *ICCV*.

Vivanco Cepeda, V.; Nayak, G. K.; and Shah, M. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *NeurIPS*.

Wang, L.; Zhang, T.; Qu, Y.; Liang, S.; Chen, Y.; Liu, A.; Liu, X.; and Tao, D. 2025. Black-Box Adversarial Attack on Vision Language Models for Autonomous Driving. *arXiv preprint arXiv:2501.13563*.

Wang, Z.; Xu, D.; Khan, R. M. S.; Lin, Y.; Fan, Z.; and Zhu, X. 2024. Llmgeo: Benchmarking large language models on image geolocation in-the-wild. *arXiv preprint arXiv:2405.20363*.

Wang, Z.; Zhang, Z.; Liang, S.; and Wang, X. 2023. Diversifying the High-level Features for better Adversarial Transferability. *arXiv preprint arXiv:2304.10136*.

Wei, X.; Liang, S.; Chen, N.; and Cao, X. 2018. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*.

Wei, Z.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2023. Enhancing the self-universality for transferable targeted attacks. In *CVPR*.

Wu, M.; and Huang, Q. 2022. Im2city: image geolocalization via multi-modal learning. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 50–61.

Xu, W.; Chen, K.; Gao, Z.; Wei, Z.; Chen, J.; and Jiang, Y.-G. 2024. Highly transferable diffusion-based unrestricted adversarial attack on pre-trained vision-language models. In *ACM MM*.

Yang, Y.; Wang, S.; Li, D.; Sun, S.; and Wu, Q. 2024. GeoLocator: A location-integrated large multimodal model (LMM) for inferring geo-privacy. *Applied Sciences*, 14(16): 7091.

Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, nwae403.

Yin, Z.; Ye, M.; Zhang, T.; Du, T.; Zhu, J.; Liu, H.; Chen, J.; Wang, T.; and Ma, F. 2023. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *NeurIPS*.

Zhang, G.; Zhang, Y.; Zhang, K.; and Tresp, V. 2024. Can vision-language models be a good guesser? exploring vlms for times and location reasoning. In *WACV*.

Zhang, J.; Ye, J.; Ma, X.; Li, Y.; Yang, Y.; Chen, Y.; Sang, J.; and Yeung, D.-Y. 2025a. AnyAttack: Towards Large-scale Self-supervised Adversarial Attacks on Vision-language Models. In *CVPR*.

Zhang, J.; Yi, Q.; and Sang, J. 2022. Towards adversarial attack on vision-language pre-training models. In *ACM MM*.

Zhang, Z.; Li, R.; Kabir, T.; and Boyd-Graber, J. 2025b. Navig: Natural language-guided analysis with vision language models for image geo-localization. *arXiv preprint arXiv:2502.14638*.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2023. On evaluating adversarial robustness of large vision-language models. *NeurIPS*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.